Neurocomputing 127 (2014) 88–97

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Determining structural identifiability of parameter learning machines

Zhi-Yong Ran*, Bao-Gang Hu

NLPR&LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 6 May 2013
Received in revised form
23 July 2013
Accepted 20 August 2013
Communicated by Shiliang Sun
Available online 25 October 2013

Keywords:
Identifiability
Parameter learning machine
Exhaustive summary
Kullback–Leibler divergence
Parameter redundancy

ABSTRACT

This paper reports an extension of our previous study on determining structural identifiability of the *generalized constraint* (*GC*) models, which are considered to be parameter learning machines. Identifiability defines a uniqueness property to the model parameters. This property is particularly important for those physically interpretable parameters in GC models. We derive identifiability criteria according to the types of models. First, by taking the models as a family of deterministic nonlinear transformations from input space to output space, we provide a criterion for examining identifiability of the *Multiple-input Multiple-output* (*MIMO*) models. This result therefore generalizes the previous one for *Single-input Single-output* (*SISO*) and *Multiple-input Single-output* (*MISO*) models. Second, if considering the models as the mean functions of input-dependent conditional distributions within stochastic framework, we derive an identifiability criterion by means of the *Kullback–Leibler divergence* (*KLD*) and regular summary. Third, time-variant models are studied based on the *exhaustive summary* method. The new identifiability criterion is valid for a range of differential/difference equation models whenever their exhaustive summaries can be obtained. Several model examples from the literature are presented to examine their identifiability property.

1. Introduction

Mathematical models have become another sensing channel for human beings to perceive, describe, and understand either *natural* or *virtual* worlds deeply. For this reason, more and more models are and will be generated for a vast variety of applications. Their modeling approaches are of course different from varied aspects. For a fast examination of the approach differences, Dubios et al. [1], Solomatine and Ostfeld [2] and Todorovski and Dzeroski [3] considered two basic modeling approaches with respect to the degree of knowledge included, namely, "*knowledge-driven*" and "*data-driven*". The knowledge-driven modeling approach is also called "*physical-based*" [2] or "*mechanistic-based*" [3] modeling approach, because the approach relies mainly on the given knowledge in modeling, such as the first principle from physics. In contrary, the data-driven modeling approach is capable of constructing a model solely from the given data without using any prior knowledge. While Todorovski and Dzeroski [3] described the application advantages and drawbacks between the two types of modeling approaches, Hu et al. [4] compared them from the viewpoints of inference methodologies (*deduction* vs. *induction*) and parameter meaning involved. Although the data-driven models have parameters for themselves, the models are considered as "*non-parametric*" because their parameters are generally unable to represent the real ones in a physical (or target) system.

In order to take advantage of each approach, a study of integrating two types of modeling approaches is reported [2–6]. Hence, "*hybrid*" models are called when the integration approach is applied to the models [5,2,3]. For stressing on a mathematical description, another term, "*generalized constraint*" (*GC*) [7,4], is adopted to call these models. Considering the large diversity and unstructured representations of prior knowledge, one can expect that the "*hybridizing*" difficulty is appeared more from imposing "*knowledge constraints*" on the models. Fig. 1 schematically depicts a GC model, which basically consists of two modules, namely, *knowledge-driven* (*KD*) submodel and *data-driven* (*DD*) submodel. For a detailed description of the GC models, one can refer [4,8,9].

Suppose a time-invariant model is considered, a general description of the GC model is given in a form of:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \theta) = \mathbf{f}_k(\mathbf{x}, \theta_k) \oplus \mathbf{f}_d(\mathbf{x}, \theta_d)$$
$$\theta = (\theta_k, \theta_d), \quad \theta_k \cap \theta_d = \varnothing \tag{1}$$

where $\mathbf{x} \in \mathscr{R}^n$ and $\mathbf{y} \in \mathscr{R}^m$ are the input and output vectors, $\mathbf{f}$ is a function for a complete model relation between $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{f}_k$ and $\mathbf{f}_d$ are the functions associated to the KD and DD submodels, respectively. $\theta \in \mathscr{R}^k$ is the parameter vector of the function $\mathbf{f}$, θ_k and θ_d are the parameter vectors associated to the functions $\mathbf{f}_k$ and $\mathbf{f}_d$ respectively. The symbol "$\oplus$" represents a coupling operation between the two submodels. Generally, the KD submodel contains physically interpretable parameters whose identifiability is of fundamental

* Corresponding author. Tel.: +86 15201064550.
E-mail addresses: zyran@nlpr.ia.ac.cn (Z.-Y. Ran), hubg@nlpr.ia.ac.cn (B.-G. Hu).

0925-2312/$ - see front matter
http://dx.doi.org/10.1016/j.neucom.2013.08.039

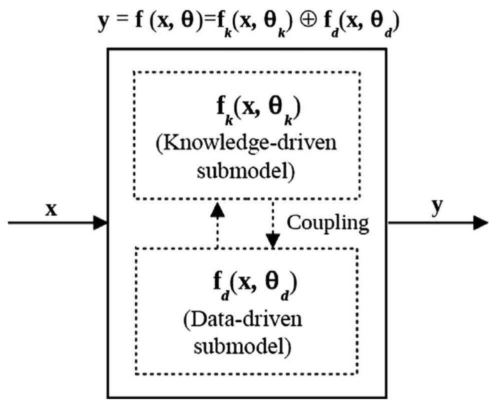

Fig. 1. Schematic diagram of GC model including KD submodel and DD submodel (modification on Figs. 1 and 3 in [4]). Two sets of parameters, θ_k and θ_d are associated with the two submodels, respectively.

importance to the understanding of the system. However, owing to the coupling operation between the two submodels, the resulting GC model may have some *unidentifiable* parameters (i.e., these parameters cannot be determined uniquely) even if the parameters of each submodels are identifiable respectively [4,8]. Identifiability of parameters will be an important aspect to reflect a transparency degree of models and hence "*determining identifiability of the models should be addressed before any implementation of estimation*" [8,10,11]. Moreover, identifiability is closely related to the convergence of a class of estimates including the maximum likelihood estimate (MLE) [8,12]. Lack of identifiability gives no guarantee of convergence to the true value of parameters and therefore usually results in severe ill-posed estimation problems [8], which is a critical issue if decisions are to be taken on the basis of their numerical values [13]. Besides the ability to detect deficient models in advance, the analysis of identifiability can also bring practical benefits, such as insightful revealing of the relations among inputs, outputs and parameters, which can be very useful for model structure design and selection [4,8]. To summarize, the usefulness and importance of identifiability analysis can be recognized in at least threefold:

(a) *Statistical inference.* In an unidentifiable statistical model, the standard statistical paradigm of the *Cramér–Rao bound* (*CRB*) does not hold, the MLE is no longer subject to Gaussian distribution even asymptotically, the model selection criteria such as AIC, BIC and MDL fail to hold, and the singularity gives rise to strange behaviors in parameter estimation, hypothesis test, Bayesian inference, model selection, etc. [14,15]. Therefore, it is imperative to check identifiability for statistical inference.
(b) *Physically interpretable* (*sub-*)*models.* In these models, some or all parameters have physically interpretable meaning [4,13,16], and to identify the true values of such parameters is important because nonuniqueness of such parameters not only means nonunique description of the process but also leads to completely erroneous or misleading results. One would not select an unidentifiable model since the parameters are of practical importance. Hence, identifiability analysis should be addressed, as part of qualitative experiment design, before any experimental data have been collected [8].
(c) *Learning dynamics.* In an unidentifiable parametric model, the trajectories of dynamics of learning are strongly affected by the nonidentifiability [14]. It has been shown that once parameters are attracted to singular points, the learning trajectory is very slow to move away from them. For example, [14] studied the dynamical behaviors of learning in multi-layer perceptions (MLP) and Gaussian mixture models (GMM), and showed that nonidentifiability resulting in plateaus and slow manifolds.

The structural identifiability is concerned with the uniqueness of the parameters determined from the input–output data. A property is said to be "*structural*" if it is true for all admissible parameter values [8]. In [4,8], the authors derived identifiability results for *Single-input Single-output* (*SISO*) and *Multiple-input Single-output* (*MISO*) models. However, their theorems cannot deal with *Multiple-input Multiple-output* (*MIMO*) models. Therefore, this work is an extension of [4,8] and we further expect to consider the problem from a wide spectrum of models. In this study, we view a model to be a "*parameter learning machine*" if it can be parameterized by a finite-dimensional vector (Fig. 2). A special emphasis is put on identifiability of arbitrary nonlinear functions for parameter learning machines. The main contribution of the present work is given from the following three aspects:

(1) From a *partial derivative matrix* (*PDM*), we derive a new identifiability criterion for deterministic nonlinear functions, which is applicable to MIMO models.
(2) Based on the *Kullback–Leibler divergence* (*KLD*) and *regular summary*, we present a new identifiability theorem for stochastic models which can be applied to more generic statistical models without restricting to exponential family [17].
(3) For the time-variant models, we adopt an *exhaustive summary* method which is valid for a wide range of differential/difference equation models whenever their exhaustive summaries can be obtained.

The remainder of this paper is organized as follows. Section 2 gives some basic definitions and views the identifiability problem from two different perspectives. Section 3 presents an identifiability criterion for deterministic MIMO models. In Section 4, we present an identifiability result for stochastic models with the help of KLD and regular summary. Section 5 gives a method for testing parameter redundancy by using exhaustive summary. Section 6 concludes with a brief summary.

2. Models and definitions

Typically, the approaches of examining structural identifiability of parameter learning machines can be categorized into two frameworks according to the modeling nature:

(1) *Deterministic framework.* In this framework, it is assumed that the model is deterministic and noise-free [8,13,16]. In other words, the model is viewed as a family of parameterized nonlinear mappings from an input vector $\mathbf{x} \in \mathscr{R}^n$ to an output vector $\mathbf{y} \in \mathscr{R}^m$,

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \theta), \tag{2}$$

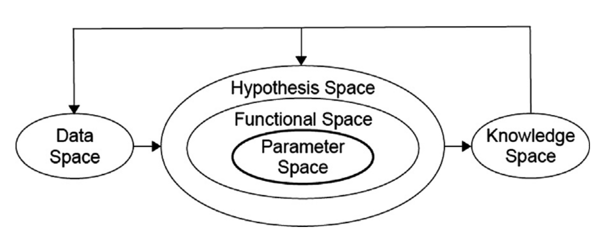

Fig. 2. Schematic diagram of spaces studied in machine learning, from which a model can be viewed as a parameter learning machine.

Table 1
General methods for testing identifiability of parametric models.

Framework	Model	Method
Deterministic	Nonlinear regression Dynamic model	Derivative function vector (DFV) method [8] Transfer function method [22] Taylor series method [23] Generating series method [24] Similarity transformation method [25] Differential algebra method [26] Implicit function theorem method [27]
Stochastic	Gaussian distribution Exponential family General distribution	Holomorphic function method [28] Derivative matrix (DM) method [17] Fisher information matrix (FIM) method [19] Kullback–Leibler divergence (KLD) method [20] Sufficient statistic method [29]

where $\theta \in \Theta$ is a parameter vector indexing a specific mapping

$$\mathscr{M}(\theta) : \theta \to \mathbf{f}(\mathbf{x}, \theta), \tag{3}$$

and $\Theta \subseteq \mathscr{R}^k$ is the admissible parameter space. In this context, structural identifiability analysis deals with the theoretic uniqueness of solutions of model parameters from perfect model specification and noise-free input–output data [8,16].

(2) *Stochastic framework*. In this framework, we introduce random noise in input and output spaces. Formally, we assume that the available data are contaminated and are generated by some stochastic system. Therefore, we can give the model a probabilistic interpretation [8,14,15]. More specifically, we assume that, given an input vector $\mathbf{x} \in \mathscr{R}^n$, the model emits an output vector $\mathbf{f}(\mathbf{x}, \theta) \in \mathscr{R}^m$ which is disturbed by a random noise $\boldsymbol{\epsilon} \in \mathscr{R}^m$. The final output $\mathbf{y} \in \mathscr{R}^m$ is

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \theta) + \boldsymbol{\epsilon}, \tag{4}$$

hence, we can interpret $\mathbf{f}(\mathbf{x}, \theta)$ as the mean function of $\mathbf{y}$ which has an input-dependent conditional distribution $p(\mathbf{y}|\mathbf{x}, \theta)$. That is, $\mathbb{E}_\theta(\mathbf{y}|\mathbf{x}, \theta) = \mathbf{f}(\mathbf{x}, \theta)$. Let $p(\mathbf{x})$ be the probability density function (PDF) over the input space $\mathscr{R}^n$, thus the joint PDF of $\mathbf{x}$ and $\mathbf{y}$ is

$$p(\mathbf{z}, \theta) = p(\mathbf{x}, \mathbf{y}, \theta) = p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}, \theta) \tag{5}$$

where $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathscr{R}^{n+m}$. Each $\theta \in \Theta$ therefore defines a PDF $p(\mathbf{z}, \theta)$ in $\mathscr{R}^{n+m}$ and we denote the corresponding probability measure by $\mathscr{M}(\theta)$.

Following [8,12,13,16], we give a unified definition for the two frameworks:

Definition 1. A model $\mathscr{M}(\theta)$, $\theta \in \Theta$ is *globally identifiable* if

$$\mathscr{M}(\theta_1) = \mathscr{M}(\theta_2) \Rightarrow \theta_1 = \theta_2, \quad \forall \theta_1, \quad \theta_2 \in \Theta. \tag{6}$$

A model is *locally identifiable* if for every $\theta \in \Theta$, there exists an open neighborhood $N(\theta)$ of θ such that the following holds

$$\mathscr{M}(\theta_1) = \mathscr{M}(\theta_2) \Rightarrow \theta_1 = \theta_2, \quad \forall \theta_1, \quad \theta_2 \in N(\theta). \tag{7}$$

Obviously, global identifiability implies local identifiability. When a parameter point $\theta_0 \in \Theta$ is of particular interest, for example, θ_0 is assumed to be the real value for the model parameter, we give the following definition.

Definition 2. A parameter point $\theta_0 \in \Theta$ is *globally identifiable* if

$$\mathscr{M}(\theta) = \mathscr{M}(\theta_0) \Rightarrow \theta = \theta_0, \quad \forall \theta \in \Theta. \tag{8}$$

A parameter point $\theta_0 \in \Theta$ is said to be *locally identifiable* if there exists an open neighborhood $N(\theta_0)$ of θ_0 such that the following holds

$$\mathscr{M}(\theta) = \mathscr{M}(\theta_0) \Rightarrow \theta = \theta_0, \quad \forall \theta \in N(\theta_0) \tag{9}$$

Remark. From Definitions 1 and 2 we can see that structural identifiability is a *theoretic* property of the model and that the presence or absence of identifiability is a feature of the specification adopted for the model, and so, is *independent* of the inferential procedure to be used [18,19]. In other words, if a model is structurally unidentifiable, no matter how carefully we design the experiment or how good the observations are, one will definitely fail to get a reasonable estimation, even when a model selection criterion (e.g., AIC, BIC, etc.) or regularization term is employed to panelize the complexity of the model [8]. Therefore, once a model has been chosen, one should test the identifiability so as to rule out prior unidentifiable models to avoid potential defects [8,16].

In this paper, special emphasis is put on *nonlinear* models which are nonlinear functions of their parameters. This is the rule for most knowledge-driven models. The structural identifiability analysis of linear models is well understood and there are a number of methods to perform such a task. When the model output is linear with respect to the parameters, the notions of local and global identifiability become equivalent, and the test for identifiability boils down to a rank condition on a data design matrix [20,21]. However, checking the identifiability is very difficult for nonlinear models. To the best of our knowledge, there are only a few methods for testing identifiability of nonlinear models. Table 1 lists the commonly used methods for checking identifiability together with their associated parametric models.

3. Identifiability criterion for deterministic models

In the deterministic framework, a model is identifiable if there exists a unique input–output behavior for each admissible parameter [8,16]. A nonlinear model that attempts to accurately describe the underlying phenomena may be complex with too many parameters. For example, a pair of parameters may always appear together as a product (or a sum) in the model equations, making it impossible to obtain unique estimate of both parameters. An open problem in nonlinear regression is to determine when different regression functions having different parameters implement identical input–output transformation [21]. In the study of machine learning, a vast majority of research has been done within the context of artificial neural networks (ANNs). For instance, for a three-layer network with H hidden units having "*tanh*" activation functions and full connectivity in both layers, there will have an overall weight space symmetry factor of $H!2^H$ [30]. In [8], Yang et al. studied structural identifiability of SISO and

MISO GC models, but their results cannot be applied to MIMO models. Generally, identifiability of deterministic nonlinear models is difficult to test since, whatever the method used, e.g., transfer function [22], generating series expansion [24], similarity transformation approach [25], differential algebra [26], implicit function theorem [27,31], it requires to solve a system of nonlinear algebraic equations whose complexity increases very fast with the number of unknown parameters, the number of input–output variables, the degree of nonlinearity of the model order, etc. Hence, it is only workable for some specific families of parametric models (e.g., polynomial and rational equations [32]) and cannot deal with arbitrary nonlinear models. To date, the precise conditions under which the input–output transformation implemented by an arbitrary nonlinear MIMO model can be uniquely determined by its parameters is a fundamental theoretical problem that has not been solved completely.

In this section, we focus our study on MIMO models within the deterministic framework. The main objective of this section involves the derivation of conditions under which a given nonlinear MIMO model will be globally identifiable.

Suppose that the MIMO model is formulated by a nonlinear vector-valued mapping $\mathbf{y}=\mathbf{f}(\mathbf{x},\theta)$ which has m component functions $f_i(\mathbf{x},\theta),\ 1\le i\le m$, more explicitly,

$$y_i=f_i(\mathbf{x},\theta)=f_i(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k),\quad 1\le i\le m. \tag{10}$$

If two parameter points θ_1 and θ_2 in Θ determine the same model, we say θ_1 is *equivalent* to θ_2, and denote $\theta_1\sim\theta_2$. That is, $\theta_1\sim\theta_2\Leftrightarrow\mathscr{M}(\theta_1)=\mathscr{M}(\theta_2)$. Note that the relation "$\sim$" is a proper equivalent relation (reflectivity, symmetry and transitivity) [20,33]. For $\theta_0\in\Theta$, we denote the *equivalence class* corresponding to θ_0 by $[\theta_0]=\{\theta\in\Theta:\theta\sim\theta_0\}$. We now present a theorem offering necessary and sufficient conditions of global identifiability for MIMO models. Our result thus generalizes the SISO and MISO results in [4,8].

Theorem 1. *(Examination of parameter identifiability for MIMO models). Suppose that an MIMO deterministic nonlinear model, denoted by $\mathbf{y}=\mathbf{f}(\mathbf{x},\theta),\quad \theta\in\Theta\subseteq\mathscr{R}^k$, is differentiable with respect to θ, and that for each $\theta\in\Theta$, $[\theta]$ is a smooth manifold of $\mathscr{R}^k$, then the model is globally identifiable if and only if the partial derivative matrix (PDM), $D=(\partial f_i/\partial\theta_j)_{m\times k}$, of $\mathbf{f}$ is symbolic column full rank, i.e., if and only if $\mathbf{v}=0$ is the unique solution of the equation $\mathbf{Dv}=0$ for all $\mathbf{x}$. In other words, the model is not globally identifiable if and only if there exists a nonzero vector $\mathbf{v}(\theta)=(v_1(\theta),\ldots,v_k(\theta))^{\mathrm{T}}$ such that the following equation holds:*

$$v_1(\theta)\frac{\partial\mathbf{f}(\mathbf{x},\theta)}{\partial\theta_1}+\cdots+v_k(\theta)\frac{\partial\mathbf{f}(\mathbf{x},\theta)}{\partial\theta_k}=0, \tag{11}$$

where the vector-valued function $\partial\mathbf{f}(\mathbf{x},\theta)/\partial\theta_i$ is defined as

$$\frac{\partial\mathbf{f}(\mathbf{x},\theta)}{\partial\theta_i}=\left(\frac{\partial f_1(\mathbf{x},\theta)}{\partial\theta_i},\cdots,\frac{\partial f_m(\mathbf{x},\theta)}{\partial\theta_i}\right)^{\mathrm{T}}. \tag{12}$$

Proof. (1) For sufficiency. If the MIMO model is not globally identifiable, then there must exist two distinct parameters $\theta_0\ne\theta_1$ in Θ, such that

$$f_i(\mathbf{x},\theta_0)=f_i(\mathbf{x},\theta_1),\quad \mathbf{x}\in\mathscr{R}^n,\quad 1\le i\le m. \tag{13}$$

Define a differentiable curve Γ as follows:

$$\Gamma=\{\theta(s)\in[\theta_0]:\theta(0)=\theta_0,\theta(1)=\theta_1,0\le s\le 1\}. \tag{14}$$

Note that the curve Γ does exist by our assumption since $[\theta_0]$ is a smooth manifold of $\mathscr{R}^k$, then $y_i,\ 1\le i\le m$ are unchanged along Γ, that is,

$$f_i(\mathbf{x},\theta(s))=const,\quad 0\le s\le 1,\quad 1\le i\le m. \tag{15}$$

Taking derivative with respect to s for each equation, we have

$$\sum_{j=1}^{k}\frac{\partial f_i}{\partial\theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}s}=0,\quad 0\le s\le 1,\quad 1\le i\le m. \tag{16}$$

That is $\mathbf{Dv}=0$ by letting $\mathbf{D}=(\partial f_i/\partial\theta_j)_{m\times k}$ and $\mathbf{v}(\theta)=(\mathrm{d}\theta_j(s)/\mathrm{d}s)_{k\times 1}$, where each $v_j(\theta)$ is independent of $\mathbf{x}$.

(2) For necessity. If there exists a non-zero vector $\mathbf{v}(\theta)=(v_1(\theta),\cdots,v_k(\theta))^{\mathrm{T}}$ such that $\mathbf{Dv}=0$, that is

$$\sum_{j=1}^{k}v_j(\theta)\frac{\partial f_i}{\partial\theta_j}=0,\quad \mathbf{x}\in\mathscr{R}^n,\quad 1\le i\le m. \tag{17}$$

This is a Lagrange linear first-order partial differential equation [34], whose auxiliary equation

$$\frac{\mathrm{d}\theta_1}{v_1(\theta)}=\cdots=\frac{\mathrm{d}\theta_k}{v_k(\theta)} \tag{18}$$

will in general have $k-1$ solutions given implicitly by, say, $a_j(\theta)=const$ for $1\le j\le k-1$. The general solution of Eq. (17) is then $f_i=h_i(a_1(\theta),\ldots,a_{k-1}(\theta))$, where h_i is an arbitrary differentiable function. Thus the model can be expressed by a smaller parameter set $\beta_j,\ 1\le j\le k-1$ by letting $\beta_j=a_j(\theta),\ 1\le j\le k-1$. This implies that the mapping $\theta\to\mathscr{M}(\theta)$ cannot be one-to-one. Therefore, the model is not globally identifiable. □

For a geometric interpretation of Theorem 1, we rewrite equation $\mathbf{Dv}=0$ as the following m equations

$$\nabla f_i^T\mathbf{v}=0,\quad i=1,\ \ldots,\ m, \tag{19}$$

where $\nabla f_1^T,\ldots,\nabla f_m^T$ are the transpose of the gradient vectors of functions $f_1,\ldots,f_m$. Each f_i is unvaried along $\mathbf{v}$ since the gradient ∇f_i of each component f_i is orthogonal to the vector field $\mathbf{v}=(v_1,\ldots,v_k)$ in the parameter space. In other words, each f_i has completely flat ridge along every smooth manifold $[\theta]$ of $\mathscr{R}^k$.

We now give some examples to illustrate the applications of Theorem 1 in examining parameter identifiability in the deterministic framework.

Example 1. (Adapted from [21]). We consider a two-input two-output deterministic model given by

$$\begin{cases} f_1(\mathbf{x},\theta)=e^{-\theta_2\theta_3x_1}+\frac{\theta_1}{\theta_2}(1-e^{-\theta_2\theta_3x_1})x_2\\ f_2(\mathbf{x},\theta)=\theta_1\theta_3x_1+(1-\theta_2\theta_3)x_2\end{cases} \tag{20}$$

with $\theta\in\mathscr{R}^3$. It can be verified that for any $\lambda\ne 0$, $(\theta_1,\theta_2,\theta_3)\sim(\lambda\theta_1,\lambda\theta_2,\theta_3/\lambda)$. Geometrically, the input–output mapping is unchanged along the differentiable curve (1-dimensional smooth manifold)

$$\Gamma=\{(\theta_1,\theta_2,\theta_3):\theta_1=t,\theta_2=t,\theta_3=1/t,t\ne 0\}. \tag{21}$$

hence, the model is not globally identifiable. We then apply Theorem 1 to this model and have

$$\mathbf{D}=\begin{pmatrix}\frac{(1-e^{-\theta_2\theta_3x_1})}{\theta_2}x_2 & \left(\frac{\theta_1\theta_3}{\theta_2}x_1x_2-\theta_3x_1-\frac{\theta_1}{\theta_2^2}x_2\right)e^{-\theta_2\theta_3x_1}-\frac{\theta_1}{\theta_2^2}x_2 & (\theta_1x_1x_2-\theta_2x_1)e^{-\theta_2\theta_3x_1}\\ \theta_3x_1 & -\theta_3x_2 & \theta_1x_1-\theta_2x_2\end{pmatrix}. \tag{22}$$

It is obvious that $\mathbf{Dv}=0$ for all $(x_1,x_2,x_3)\in\mathscr{R}^3$, where $\mathbf{v}=(\theta_1,\theta_2,-\theta_3)$, and therefore the model is not globally identifiable by Theorem 1. This verifies the validity of Theorem 1.

Example 2. [35]. Consider an MISO regression model

$$f(\mathbf{x},\theta)=\sum_{i=1}^{k}\theta_i\varphi_i(\mathbf{x}), \tag{23}$$

where $\varphi_i(\mathbf{x})$, $i=1, \ldots, k$ are known as basic functions or feature maps and $\theta \in \mathscr{R}^k$. For this type of model, we have the PDM as

$$\mathbf{D}=\left(\frac{\partial f}{\partial \theta_i}\right)_{1\times k}=(\varphi_1(\mathbf{x}),\ldots,\varphi_k(\mathbf{x})). \tag{24}$$

By Theorem 1, the model is not globally identifiable if and only if the equation

$$\mathbf{D}\mathbf{v}=\sum_{i=1}^{k} v_i\varphi_i(\mathbf{x})=0 \tag{25}$$

has nonzero solution $\mathbf{v}$. That is, $\varphi_1(\mathbf{x}),\ldots,\varphi_k(\mathbf{x})$ are functionally dependent. The validity of Theorem 1 is consistent with our intuition.

Example 3. Consider a two-input two-output nonlinear deterministic model

$$\begin{cases} y_1 = abx_1 + cdx_2 \\ y_2 = e^{-bx_1} + a\ \sin(dx_2) \end{cases} \tag{26}$$

with $\theta=(a,b,c,d)$ and $\Theta=\mathscr{R}^4$. First, we directly show that the model is globally identifiable. Otherwise, there must exist two different parameters $\theta_1=(a_1,b_1,c_1,d_1)$ and $\theta_2=(a_2,b_2,c_2,d_2)$ such that $\mathbf{f}(\mathbf{x},\theta_1)=\mathbf{f}(\mathbf{x},\theta_2)$ for all $\mathbf{x}\in\mathscr{R}^2$, that is,

$$\begin{cases} a_1b_1x_1 + c_1d_1x_2 = a_2b_2x_1 + c_2d_2x_2 \\ e^{-b_1x_1} + a_1\ \sin(d_1x_2) = e^{-b_2x_1} + a_2\ \sin(d_2x_2) \end{cases}. \tag{27}$$

From Example 2 we can see that x, e^x, $\sin x$ are functionally independent. We then have

$$a_1b_1=a_2b_2\ ,\ c_1d_1=c_2d_2,\ \ a_1\ \sin\ d_1=a_2\ \sin\ d_2,\ \ e^{-b_1}=e^{-b_2} \tag{28}$$

The above equations imply that $\theta_1=\theta_2$. This is controversial to the assumption that $\theta_1\neq\theta_2$. Hence, the model is globally identifiable. We then apply Theorem 1 to this model and have the PDM

$$\mathbf{D}=\left(\frac{\partial f_i}{\partial \theta_j}\right)_{2\times 4}=\begin{pmatrix} bx_1 & ax_1 & dx_2 & cx_2 \\ \sin(dx_2) & -x_1e^{-bx_1} & 0 & ax_2\ \cos(dx_2) \end{pmatrix} \tag{29}$$

Suppose there exists a vector $\mathbf{v}=(v_1,v_2,v_3,v_4)^T$ such that $\mathbf{D}\mathbf{v}=0$ for all $\mathbf{x}\in\mathscr{R}^4$, we will prove that $\mathbf{v}$ must be trivial. By setting $\mathbf{x}=(1,0)$, $(b,0)$, $(0,\pi/d)$, $(0,1)$, respectively, we have $v_2=0$, $v_1=0$, $v_4=0$, $v_3=0$, correspondingly. That is, the unique solution of $\mathbf{D}\mathbf{v}=0$ is $\mathbf{v}=0$. Therefore, the model is globally identifiable.

4. Identifiability criterion for stochastic models

Identifiability is a primary assumption in all classical statistical models [15,20,33]. However, such an assumption may be violated in a large variety of models. Unidentifiable families of probability distributions occur in many statistical modeling fields. In particular, in the study of machine learning, almost all learning machines used in information processing are unidentifiable [15]. Generally, if a model has hierarchical structures, latent variables or coupled submodels, the model must be unidentifiable [15].

The identifiability problem in stochastic framework is concerned with the possibility of drawing inferences from an underlying theoretical distribution. In [19], Rothenberg proved that the local identifiability of a stochastic model $p(\mathbf{z},\theta)$ is equivalent to singularity of its Fisher information matrix (FIM), i.e.,

$$\mathrm{FIM}(\theta)=-\mathbb{E}_\theta\left[\frac{\partial^2 \log\ p(\mathbf{z},\theta)}{\partial\theta^2}\right]. \tag{30}$$

A statistical learning machine is called *singular* if its FIM is singular [15,36]. The FIM is an important tool in singular learning theory, for more details about singular learning machines, one can refer [36,37]. As a special case, Hochwald et al. [28] proposed a method to establish identifiability and information-regularity of parameters in Gaussian distributions with the help of holomorphic functions. In [33], Dasgupta et al. proposed an analytical method for constructing new parameters under which an unidentifiable model will be at least locally identifiable.

Most of the previous work on identifiability problem concerned mainly with local identifiability. Up to now, few investigations have been reported on how to examine global identifiability of the models. However, as for the nonlinear regression models, we are more interested in global identifiability rather than simply local identifiability [8]. Unfortunately it is very difficult to obtain global results in a general nonlinear setting. In [19], Rothenberg established a criterion to test global identifiability for exponential family of stochastic models. Outside the exponential family it does not seem possible to get necessary and sufficient conditions for global identifiability using only the FIM. In this section, we present an applicable criterion of testing global identifiability in the stochastic framework. Essentially, non-identifiability is the consequence of the lack of enough "information" to discriminate among alternative parameter values in the model specification. Hence, it is natural to test identifiability with the help of the *Kullback–Leibler divergence* (*KLD*), which is defined as [38]

$$KL(\theta_0,\theta)=\int p(\mathbf{z},\theta_0)\log\ \frac{p(\mathbf{z},\theta_0)}{p(\mathbf{z},\theta)}d\mathbf{z}. \tag{31}$$

The KLD $KL(\theta_0,\theta)$ is always non-negative and is zero if and only if $p(\mathbf{z},\theta_0)=p(\mathbf{z},\theta)$ for every $\mathbf{z}$ [38]. To proceed in examining identifiability of parameter learning machines, a common criterion for global (local) identifiability is stated as follows [20,39].

Theorem 2. *In a stochastic model $p(\mathbf{z},\theta)$, $\theta\in\Theta$, a parameter point $\theta_0\in\Theta$ is globally (locally) identifiable if and only if θ_0 is the unique solution of the equation $KL(\theta_0,\theta)=0$ in Θ (an open neighborhood of θ_0).*

The proof can be easily verified by the fact that $KL(\theta_0,\theta)=0\Leftrightarrow p(\mathbf{z},\theta_0)=p(\mathbf{z},\theta)$ for every $\mathbf{z}$ [38]. However, for many models it is not an easy task to determine all the solutions of the equation $KL(\theta_0,\theta)=0$ in a direct way [20]. To give an example, we consider the Gaussian family

$$p(\mathbf{z},\theta)=\frac{1}{(2\pi)^{m/2}(\det\Sigma_\theta)^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mu_\theta)^{\mathrm{T}}\Sigma_\theta^{-1}(\mathbf{z}-\mu_\theta)\right\}, \tag{32}$$

where μ_θ is the mean vector and Σ_θ is the covariance matrix. The KLD can be calculated as [38]

$$KL(\theta_0,\theta)=\widetilde{KL}(\theta_0,\theta)+\frac{1}{2}(\mu_\theta-\mu_{\theta_0})^{\mathrm{T}}\Sigma_{\theta_0}^{-1}(\mu_\theta-\mu_{\theta_0}) \tag{33}$$

with

$$\widetilde{KL}(\theta_0,\theta)=\frac{1}{2}\left\{\log\ \frac{\det\Sigma_{\theta_0}}{\det\Sigma_\theta}+\mathrm{Trace}(\Sigma_\theta(\Sigma_{\theta_0}^{-1}-\Sigma_\theta^{-1}))\right\}. \tag{34}$$

It is easy to see that [38]

$$KL(\theta_0,\theta)=0\Leftrightarrow\mu_\theta=\mu_{\theta_0},\quad \Sigma_\theta=\Sigma_{\theta_0}. \tag{35}$$

Checking the identifiability of θ_0 requires us to solve a system of $m+m(m+1)/2$ nonlinear equations which makes the task intractable. Therefore, it is imperative to investigate some effective and efficient approaches to attack this problem. First we propose the following lemma.

Lemma 1. *Suppose that the parameter space Θ of a stochastic model $p(\mathbf{z},\theta)$ is a convex subset of $\mathscr{R}^k$ and that the Hessian matrix*

$$\mathbf{H}(\theta)=\left(\frac{\partial^2 KL(\theta_0,\theta)}{\partial\theta_i\partial\theta_j}\right)_{k\times k} \tag{36}$$

of the KLD $KL(\theta_0,\theta)$ is positive definite for each $\theta\in\Theta$, $\theta\neq\theta_0$, then θ_0 is globally identifiable.

Proof. It is easy to see that [38]

$$KL(\theta_0,\theta)\big|_{\theta=\theta_0}=0 \tag{37}$$

Since $\int p(\mathbf{z},\theta)\mathrm{d}\mathbf{z}=1,\quad \forall\theta\in\Theta$, we have

$$\frac{\partial\int p(\mathbf{z},\theta)\mathrm{d}\mathbf{z}}{\partial\theta}\Big|_{\theta=\theta_0}=\frac{\partial 1}{\partial\theta}\Big|_{\theta=\theta_0}=0. \tag{38}$$

hence, by interchange of integral and derivative, we get

$$\begin{aligned}\frac{\partial KL(\theta_0,\theta)}{\partial\theta}\Big|_{\theta=\theta_0}&=\frac{\partial(\int p(\mathbf{z},\theta_0)\log(p(\mathbf{z},\theta_0)/p(\mathbf{z},\theta))\mathrm{d}\mathbf{z})}{\partial\theta}\Big|_{\theta=\theta_0}\\&=\int\left(\frac{\partial p(\mathbf{z},\theta_0)\log(p(\mathbf{z},\theta_0)/p(\mathbf{z},\theta))}{\partial\theta}\right)\Big|_{\theta=\theta_0}\mathrm{d}\mathbf{z}\\&=-\int\left(\frac{\partial p(\mathbf{z},\theta)}{\partial\theta}\right)\Big|_{\theta=\theta_0}\mathrm{d}\mathbf{z}\\&=-\frac{\partial\int p(\mathbf{z},\theta)\mathrm{d}\mathbf{z}}{\partial\theta}\Big|_{\theta=\theta_0}=0.\end{aligned} \tag{39}$$

Apply *Taylor's formula* to $KL(\theta_0,\theta)$, we have

$$\begin{aligned}KL(\theta_0,\theta)&=KL(\theta_0,\theta)|_{\theta=\theta_0}+(\theta-\theta_0)^{\mathrm{T}}\left(\frac{\partial KL(\theta_0,\theta)}{\partial\theta}\right)\Big|_{\theta=\theta_0}\\&\quad+\frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}\mathbf{H}(\theta^*)(\theta-\theta_0),\end{aligned} \tag{40}$$

where

$$\mathbf{H}(\theta^*)=\frac{\partial^2 KL(\theta_0,\theta)}{\partial\theta^2}\Big|_{\theta=\theta^*},\quad \theta^*=(1-t)\theta_0+t\theta,\ 0<t<1 \tag{41}$$

hence,

$$KL(\theta_0,\theta)=\frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}\mathbf{H}(\theta^*)(\theta-\theta_0). \tag{42}$$

Since $\theta^*\neq\theta_0$, $\mathbf{H}(\theta^*)$ is positive definite. Hence

$$KL(\theta_0,\theta)>0 \text{ for any } \theta\neq\theta_0 \tag{43}$$

That is, θ_0 is the unique solution of the equation $KL(\theta_0,\theta)=0$. By Theorem 2, θ_0 is globally identifiable. □

In order to provide some efficient and applicable criteria, we should resort to two key quantities, namely the *exhaustive summary* and *regular summary*, which can help to determine the parameter structure of the model. An exhaustive summary is a vector-valued function of original parameters that uniquely defines the model, and a formal definition is given below, adapted from [24].

Definition 3. A vector-valued function $\mathbf{s}(\theta)=(s_1(\theta),\ldots,s_q(\theta))^T$, is an *exhaustive summary* if each $s_i(\theta)$, $i=1,\ \ldots,\ q$ is a non-constant function and the mapping $\mathbf{s}(\theta)\to\mathscr{M}(\theta)$ is bijective. That is, the following condition holds:

$$\mathscr{M}(\theta_1)=\mathscr{M}(\theta_2)\Leftrightarrow\mathbf{s}(\theta_1)=\mathbf{s}(\theta_2),\quad\forall\theta_1,\ \theta_2\in\Theta. \tag{44}$$

A vector-valued function $\mathbf{s}(\theta)=(s_1(\theta),\ldots,s_q(\theta))^T$ of θ is a *regular summary* if $\mathbf{H}(\mathbf{s})$ is positive definite for all $\mathbf{s}$, where $\mathbf{H}(\mathbf{s})$ is the Hessian matrix of $KL(\mathbf{s}_0,\mathbf{s})$.

In the above definition, we make the assumption that each $s_i(\theta)$ is not a constant function, as a constant component in $\mathbf{s}(\theta)$ is helpless in determining the parameter structure of $\mathscr{M}(\theta)$. Moreover, Eq. (44) ensures that the mapping $\mathbf{s}(\theta)\to\mathscr{M}(\theta)$ cannot be trivial. Take the Gaussian model (Eq. (32)) as an example, the exhaustive summary is formed from the non-constant elements in the $m\times 1$ mean vector μ_θ and the $m(m+1)/2$ non-constant, non-duplicated elements in the covariance matrix Σ_θ (See Example 4).

We then give an identifiability result for stochastic models with the help of KLD and regular summary.

Theorem 4. *Suppose that $p(\mathbf{z},\theta)$, $\theta\in\Theta$ is a stochastic model and that $\mathbf{s}(\theta)$ is a regular summary, if the Jacobian matrix $\mathbf{J}(\theta)=(\partial\mathbf{s}/\partial\theta)=(\partial s_i/\partial\theta_j)$ is of symbolic column full rank, i.e., $\mathbf{J}(\theta)$ is of full rank for all $\theta\in\Theta$, then the model $p(\mathbf{z},\theta)$, $\theta\in\Theta$ is globally identifiable.*

Proof. Since $\int p(\mathbf{z},\theta)\mathrm{d}\mathbf{z}=1,\quad\forall\theta\in\Theta$, we have

$$\begin{aligned}\mathbb{E}_\theta\left(\frac{1}{p(\mathbf{z},\theta)}\frac{\partial^2 p(\mathbf{z},\theta)}{\partial\theta_i\partial\theta_j}\right)&=\int\frac{\partial^2 p(\mathbf{z},\theta)}{\partial\theta_i\partial\theta_j}\mathrm{d}\mathbf{z}\\&=\frac{\partial^2\int p(\mathbf{z},\theta)\mathrm{d}\mathbf{z}}{\partial\theta_i\partial\theta_j}=\frac{\partial^2 1}{\partial\theta_i\partial\theta_j}=0\end{aligned} \tag{45}$$

By simple calculation we get

$$\frac{\partial^2\log\ p(\mathbf{z},\theta)}{\partial\theta_i\partial\theta_j}=-\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta_i}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta_j}+\frac{1}{p(\mathbf{z},\theta)}\frac{\partial^2 p(\mathbf{z},\theta)}{\partial\theta_i\partial\theta_j}. \tag{46}$$

hence

$$\mathbb{E}_\theta\left(\frac{\partial^2\log\ p(\mathbf{z},\theta)}{\partial\theta_i\partial\theta_j}\right)=-\mathbb{E}_\theta\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta_i}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta_j}\right). \tag{47}$$

That is,

$$\mathbb{E}_\theta\left(\frac{\partial^2\log\ p(\mathbf{z},\theta)}{\partial\theta^2}\right)=-\mathbb{E}_\theta\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta^{\mathrm{T}}}\right). \tag{48}$$

For $\theta_0\in\Theta$, by interchange of integral and derivative, we have

$$\begin{aligned}\mathbf{H}(\theta_0)=\frac{\partial^2 KL(\theta_0,\theta)}{\partial\theta^2}\Big|_{\theta=\theta_0}&=-\frac{\partial^2\int p(\mathbf{z},\theta_0)\log\ p(\mathbf{z},\theta)\mathrm{d}\mathbf{z}}{\partial\theta^2}\Big|_{\theta=\theta_0}\\&=-\left(\int\frac{\partial^2(p(\mathbf{z},\theta_0)\log\ p(\mathbf{z},\theta))}{\partial\theta^2}\mathrm{d}\mathbf{z}\right)\Big|_{\theta=\theta_0}\\&=-\int p(\mathbf{z},\theta_0)\left(\frac{\partial^2\log\ p(\mathbf{z},\theta)}{\partial\theta^2}\right)\Big|_{\theta=\theta_0}\mathrm{d}\mathbf{z}\\&=-\mathbb{E}_{\theta_0}\left(\frac{\partial^2\log\ p(\mathbf{z},\theta)}{\partial\theta^2}\right)\\&=\mathbb{E}_{\theta_0}\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta^{\mathrm{T}}}\right).\end{aligned} \tag{49}$$

From Eq. (38) we have

$$\begin{aligned}\mathbb{E}_{\theta_0}\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta}\right)&=\int\left(p(\mathbf{z},\theta)\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta}\right)\Big|_{\theta=\theta_0}\mathrm{d}\mathbf{z}\\&=\int\left(\frac{\partial p(\mathbf{z},\theta)}{\partial\theta}\right)\Big|_{\theta=\theta_0}\mathrm{d}\mathbf{z}=0\end{aligned} \tag{50}$$

hence

$$\mathbf{H}(\theta_0)=\left(\mathrm{Cov}\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta}\right)\right)\Big|_{\theta=\theta_0}, \tag{51}$$

where $(\mathrm{Cov}(\partial\log\ p(\mathbf{z},\theta)/\partial\theta))|_{\theta=\theta_0}$ is the covariance matrix of the random vector $\partial\log\ p(\mathbf{z},\theta)/\partial\theta$ evaluated at θ_0. Denote $\mathbf{H}(\theta)=(H_{ab}(\theta))$. From Eq. (49) we have

$$\begin{aligned}H_{ab}(\theta)&=\mathbb{E}_\theta\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta_a}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial\theta_b}\right)\\&=\mathbb{E}_\theta\left(\left(\sum_{i=1}^{q}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial s_i}\frac{\partial s_i}{\partial\theta_a}\right)\left(\sum_{j=1}^{q}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial s_j}\frac{\partial s_j}{\partial\theta_b}\right)\right)\\&=\sum_{i,j=1}^{q}\mathbb{E}_\theta\left(\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial s_i}\frac{\partial\log\ p(\mathbf{z},\theta)}{\partial s_j}\right)\frac{\partial s_i}{\partial\theta_a}\frac{\partial s_j}{\partial\theta_b}\\&=\sum_{i,j=1}^{q}H_{ab}(\mathbf{s})\frac{\partial s_i}{\partial\theta_a}\frac{\partial s_j}{\partial\theta_b}\end{aligned} \tag{52}$$

Rewrite the above equation in a compact form, we have

$$\mathbf{H}(\theta)=\mathbf{J}(\theta)^T\mathbf{H}(\mathbf{s})\mathbf{J}(\theta). \tag{53}$$

Since $\mathbf{s}(\theta)$ is a regular summary, $\mathbf{H}(\mathbf{s})$ is positive definite. $\mathbf{H}(\theta)$ is positive definite as $\mathbf{J}(\theta)$ is of column full-rank. Hence, θ is globally identifiable by Lemma 1. From Definition 1 we can see that a model $p(\mathbf{z},\theta)$ is globally identifiable if and only if $p(\mathbf{z},\theta)$ is globally identifiable at every $\theta\in\Theta$. Since θ is an arbitrary point in Θ, the model $p(\mathbf{z},\theta)$ is globally identifiable. □

Corollary 1. *Suppose the stochastic model $p(\mathbf{z},\theta)$ is from an exponential family*

$$p(\mathbf{z},\theta)=\xi(\theta)c(\mathbf{z})\exp\left\{\sum_{i=1}^{q}\eta_i(\theta)T_i(\mathbf{z})\right\} \tag{54}$$

and $\eta(\theta)=(\eta_1(\theta),\dots,\eta_q(\theta))^T$ is the natural parameter vector, if the Jacobian matrix $\mathbf{J}(\theta)=\partial\eta/\partial\theta$ is of symbolic column full rank, then $p(\mathbf{z},\theta)$ is globally identifiable.

Proof. Let $\mathbf{z}_1,\ \dots,\ \mathbf{z}_n$ be an independent and identically distributed (i.i.d.) sample from $p(\mathbf{z},\theta)$, $\sum_{j=1}^{n}T_i(\mathbf{z}_j)$ is the sufficient statistic of $\eta_i(\theta)$ since $\eta_i(\theta)$ is the natural parameter [40]. By the sufficient statistic method [29] we can see that $p(\mathbf{z},\eta)$ is globally identifiable. Hence, $p(\mathbf{z},\eta)$ satisfies Cramér–Rao regularity conditions [40]. Therefore, the Hessian matrix $\mathbf{H}(\eta)$ is positive definite since it is the covariance matrix of random vector $\partial\log\ p(\mathbf{z},\eta)/\partial\eta$. From Eq. (53) we have

$$\mathbf{H}(\theta)=\mathbf{J}(\theta)^{\mathrm{T}}\mathbf{H}(\eta)\mathbf{J}(\theta). \tag{55}$$

Since $\mathbf{J}(\theta)$ is of symbolic column full rank, from Theorem 4, $p(\mathbf{z},\theta)$ is globally identifiable. □

A remarkable feature of Theorem 4 and Corollary 1 is that we can determine the identifiability of stochastic models by calculating the symbolic rank of the Jacobian matrix $\mathbf{J}(\theta)$, thus avoiding the usual bottleneck of seeking for the roots of a nonlinear equation $KL(\theta_0,\theta)=0$. Our result can be applied in a variety of stochastic models without restricting to exponential family of distributions.

Example 4. [41]. Consider the second-order state-space model

$$\begin{pmatrix}x_1(t+1)\\ x_2(t+1)\end{pmatrix}=\begin{pmatrix}\theta_1 & 0\\ 1 & \theta_2\end{pmatrix}\begin{pmatrix}x_1(t)\\ x_2(t)\end{pmatrix}+\begin{pmatrix}1\\ 0\end{pmatrix}\epsilon(t),\quad \begin{pmatrix}x_1(0)\\ x_2(0)\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix},$$
$$y(t)=x_2(t) \tag{56}$$

where $\theta=(\theta_1,\theta_2)\in\mathscr{R}^2$ and the noise $\epsilon(t)$ is a zero-mean Gaussian white noise with unit power. Let us study the output sequence with $t=4$. The output sequence $\mathbf{y}^4$ is as follows:

$$\mathbf{y}^4=\begin{pmatrix}y(1)\\ y(2)\\ y(3)\\ y(4)\end{pmatrix}=\begin{pmatrix}0\\ \epsilon(0)\\ (\theta_1+\theta_2)\epsilon(0)+\epsilon(1)\\ (\theta_1^2+\theta_1\theta_2+\theta_2^2)\epsilon(0)+(\theta_1+\theta_2)\epsilon(1)+\epsilon(2)\end{pmatrix}. \tag{57}$$

It is easy to see that $\mathbf{y}^4\sim\mathcal{N}(0,\Sigma(\theta))$ is a zero-mean Gaussian vector whose distribution can be uniquely determined by its covariance matrix $\Sigma(\theta)$. Let $\mathbf{s}(\theta)$ be a vector containing all the distinct non-constant elements of $\Sigma(\theta)$, that is,

$$\mathbf{s}(\theta)=\begin{pmatrix}\theta_1+\theta_2\\ \theta_1^2+\theta_1\theta_2+\theta_2^2\\ (\theta_1+\theta_2)^2+1\\ (\theta_1+\theta_2)(\theta_1^2+\theta_1\theta_2+\theta_2^2+1)\\ (\theta_1^2+\theta_1\theta_2+\theta_2^2)^2+(\theta_1+\theta_2)^2+1\end{pmatrix}. \tag{58}$$

Obviously, $\mathbf{s}(\theta)$ is a regular summary of $\mathbf{y}^4$. The Jacobian matrix $\mathbf{J}(\theta)$ can be calculated as

$$\mathbf{J}(\theta)=\begin{pmatrix}1 & 1\\ 2\theta_1+\theta_2 & \theta_1+2\theta_2\\ 2(\theta_1+\theta_2) & 2(\theta_1+\theta_2)\\ 3\theta_1^2+4\theta_1\theta_2+2\theta_2^2+1 & 2\theta_1^2+4\theta_1\theta_2+3\theta_2^2+1\\ 2(2\theta_1^3+3\theta_1^2\theta_2+3\theta_1\theta_2^2+\theta_2^3+\theta_1+\theta_2) & 2(2\theta_1^3+3\theta_1^2\theta_2+3\theta_1\theta_2^2+\theta_2^3+\theta_1+\theta_2)\end{pmatrix}. \tag{59}$$

Further, by using elementary matrix transformation, we can see that $\mathbf{J}(\theta)$ is equivalent to

$$\begin{pmatrix}1 & 0\\ 2\theta_1+\theta_2 & -\theta_1+\theta_2\\ 2(\theta_1+\theta_2) & 0\\ 3\theta_1^2+4\theta_1\theta_2+2\theta_2^2+1 & -\theta_1^2+\theta_2^2\\ 2(2\theta_1^3+3\theta_1^2\theta_2+3\theta_1\theta_2^2+\theta_2^3+\theta_1+\theta_2) & 0\end{pmatrix} \tag{60}$$

we have rank $(\mathbf{J}(\theta))=2$ for all $\theta\in\mathscr{R}^2$ such that $\theta_1\neq\theta_2$. Hence, $\mathbf{H}(\theta)$ is positive definite for all $\theta\in\mathscr{R}^2$ such that $\theta_1\neq\theta_2$. From Corollary 1, the model is globally identifiable for all $\theta\in\mathscr{R}^2$ such that $\theta_1\neq\theta_2$. According to [41], the model is locally identifiable by their transfer function method, but our method gives a much stronger conclusion.

Example 5. Consider the 1-order autoregressive (AR) model

$$y_t=\theta_1 y_{t-1}+\theta_2\epsilon_t,\quad 0<\theta_1<1,\quad \theta_2\neq 0 \tag{61}$$

with ϵ_t a zero-mean Gaussian white noise with unit power and $\theta=(\theta_1,\theta_2)$. Assume that the system has reached steady state when the observations begin, then the observation sequence $\{y_t\}$ will be a 1-order stationary Markov process whose covariance matrix is

$$\Sigma_t(\theta)=\frac{\theta_2^2}{1-\theta_1^2}\begin{pmatrix}1 & \theta_1 & \theta_1^2 & \cdots & \theta_1^{t-1}\\ \theta_1 & 1 & \theta_1 & \cdots & \theta_1^{t-2}\\ \theta_1^2 & \theta_1 & 1 & \cdots & \theta_1^{t-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \theta_1^{t-1} & \theta_1^{t-2} & \theta_1^{t-3} & \cdots & 1\end{pmatrix}_{t\times t},\quad t=1,\ 2,\dots \tag{62}$$

Let $\mathbf{s}_t(\theta)$ be a vector containing all the distinct non-constant elements of $\Sigma_t(\theta)$, that is,

$$\mathbf{s}_t(\theta)=\frac{\theta_2^2}{1-\theta_1^2}(1,\theta_1,\dots,\theta_1^{t-1})^T. \tag{63}$$

Obviously, $\mathbf{s}_t(\theta)$ is a regular summary of the observation sequence $\{y_t\}$. The Jacobian matrix $\mathbf{J}_t(\theta)$ can be calculated as

$$\mathbf{J}_t(\theta)=\frac{\theta_2}{1-\theta_1^2}\begin{pmatrix}2\theta_1\theta_2 & 2\\ (1+\theta_1)^2\theta_2 & 2\theta_1\\ \vdots & \vdots\\ ((n-1)\theta_1^{t-2}+(n+1)\theta_1^t)\theta_2 & 2\theta_1^{t-1}\end{pmatrix}. \tag{64}$$

We have $\mathrm{rank}(\mathbf{J}_t(\theta))=2$ for all θ such that $0<\theta_1<1,\ \ \theta_2\neq 0$. From Corollary 1, the model is globally identifiable for all θ such that $0<\theta_1<1,\ \ \theta_2\neq 0$.

5. Parameter redundancy

The most obvious cause of non-identifiability is parameter redundancy, in the sense that the model can be written in terms of a smaller set of parameters. Following [8,17], we give the following definition.

Definition 4. (Parameter redundancy). A model $\mathscr{M}(\theta),\ \theta\in\Theta\subset\mathscr{R}^k$ is said to be parameter *redundant* if it can be expressed in terms of a smaller parameter vector $\beta=\beta(\theta)$, where $\dim\beta<k$. Models which are not parameter redundant are said to be of *full rank*.

In [17], Catchpole et al. introduced the concept of parameter redundancy in exponential family of distributions and they further showed that whether or not a model is parameter redundant can be determined by checking the symbolic rank of a *derivative matrix* (*DM*), but their DM-based method can only be used in the exponential case. In this section, we will extend the result for exponential family to more generic models. By using exhaustive summaries, we provide a criterion for checking identifiability of models as follows.

Theorem 6. *Suppose that* $\mathbf{s}(\theta) = (s_1(\theta), \dots, s_q(\theta))^T$ *is the exhaustive summary of the model* $\mathcal{M}(\theta)$, $\theta \in \mathcal{R}^k$, *then* $\mathcal{M}(\theta)$ *is parameter redundant if and only if the Jacobian matrix*

$$\frac{\partial \mathbf{s}}{\partial \theta} = \left(\frac{\partial s_i}{\partial \theta_j}\right)_{q \times k} \tag{65}$$

is symbolically column rank-deficient, i.e., the Jacobian matrix is column-deficient for all θ.

Proof. For necessity. Since $\mathcal{M}(\theta)$ is parameter redundant, then the exhaustive summary $\mathbf{s}(\theta)$ can be expressed by a smaller parameter vector $\beta = \beta(\theta)$, $\dim \beta = r < k$. Specifically, let $\beta = (\beta_1, \dots, \beta_r)$, we have

$$s_i(\theta_1, \dots, \theta_k) = s_i(\beta_1, \dots, \beta_r) = s_i(\beta_1(\theta_1, \dots, \theta_k), \dots, \beta_r(\theta_1, \dots, \theta_k)) \tag{66}$$

Taking derivative with respect to θ_j for each equation, we have

$$\frac{\partial s_i}{\partial \theta_j} = \sum_{l=1}^{r} \frac{\partial s_i}{\partial \beta_l}\frac{\partial \beta_l}{\partial \theta_j}, \quad i = 1, \dots, q; \quad j = 1, \dots, k. \tag{67}$$

That is

$$\begin{pmatrix} \frac{\partial s_1}{\partial \theta_1} & \cdots & \frac{\partial s_1}{\partial \theta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_q}{\partial \theta_1} & \cdots & \frac{\partial s_q}{\partial \theta_k} \end{pmatrix}_{q \times k} = \begin{pmatrix} \frac{\partial s_1}{\partial \beta_1} & \cdots & \frac{\partial s_1}{\partial \beta_r} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_q}{\partial \beta_1} & \cdots & \frac{\partial s_q}{\partial \beta_r} \end{pmatrix}_{q \times r} \begin{pmatrix} \frac{\partial \beta_1}{\partial \theta_1} & \cdots & \frac{\partial \beta_1}{\partial \theta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial \beta_r}{\partial \theta_1} & \cdots & \frac{\partial \beta_r}{\partial \theta_k} \end{pmatrix}_{r \times k}. \tag{68}$$

We rewrite the above matrix equation in a compact form

$$\left(\frac{\partial \mathbf{s}}{\partial \theta}\right)_{q \times k} = \left(\frac{\partial \mathbf{s}}{\partial \beta}\right)_{q \times r} \left(\frac{\partial \beta}{\partial \theta}\right)_{r \times k}. \tag{69}$$

It is easy to see that

$$\operatorname{rank}\left(\frac{\partial \mathbf{s}}{\partial \theta}\right)_{q \times k} \le \operatorname{rank}\left(\frac{\partial \beta}{\partial \theta}\right)_{r \times k} \le r < k. \tag{70}$$

Therefore, the Jacobian matrix $\partial \mathbf{s}/\partial \theta$ is symbolically column rank-deficient. The sufficiency can be derived in the same line as Theorem 1.

In the study of modeling dynamical systems using differential equations for which closed-form solutions are not available, parameter redundancy analysis is an important tool to study the problem of structural identifiability.

Example 7. Consider the following dynamic ordinary differential equation (ODE) model [32]:

$$\begin{cases} \dot{x}_1 = -\theta_2 x_1 - \theta_3 x_2 - \theta_0 u \\ \dot{x}_2 = -\theta_1 x_1 + \theta_3 x_1 x_2 \\ y = x_1 + \epsilon \end{cases}, \tag{71}$$

where $\theta = (\theta_0, \theta_1, \theta_2, \theta_3), \theta_i \ne 0$, $i = 0, \dots, 3$, u is the input variable, x_j, $j = 1, 2$ are the state variables, y is the output variable and ϵ is the random noise. First, we have the noisy input–output model as [32]

$$\begin{aligned} &-\ddot{y} - \ddot{\epsilon} - \theta_0 \dot{u} - \theta_2(\dot{y} + \dot{\epsilon}) + \theta_3(\dot{y} + \dot{\epsilon})(y + \epsilon) \\ &+ \theta_0\theta_3 u(y + \epsilon) + \theta_2\theta_3(y + \epsilon)^2 + \theta_1\theta_3(y + \epsilon) = 0 \end{aligned} \tag{72}$$

The exhaustive summary is $\mathbf{s}(\theta) = (\theta_0, \theta_2, \theta_3, \theta_0\theta_3, \theta_2\theta_3, \theta_1\theta_3)^T$ and the Jacobian matrix $\partial \mathbf{s}/\partial \theta$ is

$$\frac{\partial \mathbf{s}}{\partial \theta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \theta_3 & 0 & 0 & \theta_0 \\ 0 & 0 & \theta_3 & \theta_2 \\ 0 & \theta_3 & 0 & \theta_1 \end{pmatrix}. \tag{73}$$

It is easy to check that rank $(\partial \mathbf{s}/\partial \theta) = 4$ for every θ, so the model is of full rank and therefore not parameter redundant. The identifiability of the system can also be checked by differential algebra method [32]. The two approaches give the same result, but our method needs not to solve a system of nonlinear equations.

Example 8. Consider a 4-D HIV/AIDS model [27]

$$\begin{cases} \dot{T} = s - dT - \beta v T \\ \dot{T}_1 = q_1 \beta v T - \mu_1 T_1 - k_1 T_1 \\ \dot{T}_2 = q_2 \beta v T + k_1 T_1 - \mu_2 T_2 \\ \dot{v} = k_2 T_2 - cv \\ y_1(t) = T(t) \\ y_2(t) = v(t) \end{cases}. \tag{74}$$

Here the unknown parameter $\theta = (\beta, d, s, q_1, k_1, \mu_1, q_2, k_2, \mu_2, c)$ and the initial conditions of the model are assumed to be known. The main question to be addressed is whether θ is globally identifiable from an experiment in which the output functions $y_1(t), y_2(t)$ are exactly measured. The exhaustive summary $\mathbf{s}(\theta)$ is as follows [31]:

$$\mathbf{s}(\theta) = \begin{pmatrix} \beta \\ d \\ s \\ c + k_1 + \mu_1 + \mu_2 \\ \beta k_2 q_2 \\ ck_1 + c\mu_1 + c\mu_2 + k_1\mu_2 + \mu_1\mu_2 \\ \beta^2 k_2 q_2 \\ \beta k_2(dq_2 - k_1 q_1 - k_1 q_2 - \mu_1 q_2) \\ -\beta k_2 q_2 s + ck_1\mu_2 + c\mu_1\mu_2 \end{pmatrix}. \tag{75}$$

The Jacobian matrix $\partial \mathbf{s}/\partial \theta$ can be written as a 2-by-2 block matrix

$$\frac{\partial \mathbf{s}}{\partial \theta} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}, \tag{76}$$

where $\mathbf{M}_{11}$ is a 3-by-3 identity matrix. It is obvious that the first three columns of $\partial \mathbf{s}/\partial \theta$ is column independent and hence the parameters β, d, s are globally identifiable. Let $\mathbf{s}_1(\theta)$ be the sub-vector of $\mathbf{s}(\theta)$ with the terms (β, d, s) excluded and $\theta^1 = (q_1, k_1, \mu_1, q_2, k_2, \mu_2, c)$, the column vectors of the Jacobian matrix of $\mathbf{M}_{22} = \partial \mathbf{s}_1/\partial \theta^1$ are given as follows:

$$\begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ \beta k_2 \\ \beta q_2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ c + \mu_2 \\ c + \mu_2 \\ 0 \\ 0 \\ c + k_1 + \mu_1 \\ k_1 + \mu_1 + \mu_2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ \beta^2 k_2 \\ \beta^2 q_2 \\ 0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} -\beta k_1 k_2 \\ -\beta k_2(q_1+q_2) \\ -\beta k_2 q_2 \\ \beta k_2(d-k_1-\mu_1) \\ \beta(dq_2-k_1q_1-k_1q_2-\mu_1 q_2) \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ c\mu_2 \\ c\mu_2 \\ -\beta k_2 s \\ -\beta q_2 s \\ c(k_1+\mu_1) \\ \mu_2(k_1+\mu_1) \end{pmatrix}. \quad (77)$$

Since the second and the fourth column vectors of $\partial \mathbf{s}_1/\partial\theta^1$ are linearly dependent, parameter vector θ^1 is unidentifiable. Our method gives the same result as the one given by [27], but our method gives a solution within a much fewer steps.

6. Conclusion

Identifiability becomes an essential requirement for learning machines when the models contain physically interpretable parameters. Despite the existing methods can handle some specific families of parameter models, the structural identifiability analysis for arbitrary nonlinear models is still an open question [8,21]. This paper is a further study on the structural identifiability of parameter learning machines. For the time-invariant models, we first present an identifiability result for MIMO models within the deterministic framework. Our result generalizes the previous one for SISO and MISO models proposed in [4,8]. In addition, we develop an identifiability criterion by means of KLD and regular summary within the stochastic framework. The resulting theorem can be applied in a variety of distributions not restricted to exponential families. For the time-variant models, we adopt an exhaustive summary method which is valid for a wide range of differential/difference equation models whenever their exhaustive summaries can be obtained.

Finally, we outline two directions below for future work:

(1) One of the major objectives in the analysis of identifiability problem is to obtain a set of identifying functions and then use them to reparameterize the model for subsequent analysis and estimation [33]. In almost all cases, such a set of functions cannot be easily obtained by visual inspection or by simple analytic verification. In our present paper, we propose some criteria to test structural identifiability in parameter learning machines, but it tells nothing about reparameterization when parameter redundancy is detected. It is still an open problem which is one of the directions of research into the identifiability theory [33].

(2) For the time-variant models, the exhaustive summary method we adopted is theoretically general but may be not practicably applicable to any parameter models. So far the exhaustive summary method has worked in a range of *ordinary differential equation* (*ODE*) models. However, it is a hard task to obtain the exhaustive summaries in *partial differential equation* (*PDE*) models via Laplace transformation [22], Taylor series method [23], etc. Therefore, it would be highly desirable to consider the alternative methods for obtaining exhaustive summaries in PDE models.

Acknowledgments

This work is supported in part by NSFC no. 61273196.

References

[1] D. Dubios, P. Hajek, H. Prade, Knowledge-driven versus data-driven logics, J. Logics Lang. Inf. 9 (2000) 65–89.
[2] D.P. Solomatine, A. Ostfeld, Data-driven modeling: some past experiences and new approaches, J. Hydroinf. 10 (1) (2008) 3–22.
[3] L. Todorovski, S. Dzeroski, Integrating knowledge-driven and data-driven approaches to modeling, Ecol. Modeling 194 (1–3) (2006) 3–13.
[4] B.-G. Hu, H.B. Qu, S.H. Yang, A generalized-constraint neural network model: associating partially known relationships for nonlinear regressions, Inf. Sci. 179 (2009) 1929–1943.
[5] D. Psichogios, L.H. Ungar, A hybrid neural network – first principles approach to process modeling, AIChE J. 38 (1992) 1499–1511.
[6] R. Ben-Hamadou, N. Atanasova, E. Wolanski, Ecohydrology modeling: tools for management, in: E Wolanski, DS Mclusky (Eds.), Treatise on Estuarine and Coastal Science, 10, Academic Press, Waltham, 2001, pp. 301–328.
[7] L.A. Zadeh, The concept of a generalized constraint – a bridge from natural languages to mathematics, in: NAFIPS 2005-Annual Meeting of the North American Fuzzy Information Processing Society, 2005, pp. 1–2.
[8] S.H. Yang, B.-G Hu, P.H. Cournède, Structural identifiability of generalized-constraint neural network models for nonlinear regression, Neurocomputing 72 (2008) 392–400.
[9] Y.J. Qu, B.-G. Hu, Generalized constraint neural network regression model subject to linear priors, IEEE Tran. Neural Networks 22 (2011) 2447–2459.
[10] D. Csercsik, K.M. Hangos, G. Szederkenyi, Identifiability analysis and parameter estimation of a single Hodgkin–Huxley type voltage dependent ion channel under voltage step measurement conditions, Neurocomputing 77 (2012) 178–188.
[11] M. Atencia, G. Joya, F. Sandoval, Identification of noisy dynamical systems with parameter estimation based on Hopfield neural networks, Neurocomputing 121 (2013) 14–24.
[12] L. Ljung, System Estimation: Theory for the User, Second ed., Prentice-Hall, Englewood Cliffs, NJ, 1999.
[13] L. Wang, H. Garnier, System Estimation, Environmental Modeling and Control System Design, Springer, London, 2012.
[14] S.I. Amari, H. Park, T. Ozeki, Singularities affect dynamics of learning in neuromanifolds, Neural Comput. 18 (2006) 1007–1065.
[15] S.Watanabe, Almost all Learning Machines are Singular, Invited Paper in FOCI, 2007.
[16] S. Audoly, L. D'Angio, M.P. Saccomani, C. Cobelli, Global identifiability of biokinetic models of linear compartment models, a computer algebra algorithm, IEEE Trans. Biomed. Eng. 45 (1998) 36–47.
[17] E.A. Catchpole, B.J.T. Morgan, Detecting parameter redundancy, Biometrika 84 (1) (1997) 187–196.
[18] G. Casella, R.L. Berger, Statistical Inference, Second ed., Duxbury, 2002.
[19] T.J. Rothenberg, Identification in parametric models, Econometrica 39 (3) (1971) 577–591.
[20] C.D.M. Paulino, C.A.D.B. Pereira, On identifiability of parametric statistical models, J. Ital. Stat. Soc. 3 (1994) 125–151.
[21] G.A.F. Seber, C.J. Wild, Nonlinear Regression, Wiley, New York, 2003.
[22] R. Bellman, K.J. Astrom, On structural identifiability, Math. Biosci. 7 (1970) 329–339.
[23] H. Pohjanpalo, System identifiability based on power-series expansion of solution, Math. Biosci. 41 (1978) 21–33.
[24] E. Walter, Y. Lecourtier, Global approaches to identifiability testing for linear and nonlinear state space models, Math. Comput. Simulation 24 (1982) 472–482.
[25] S. Vajda, K. Godfrey, H. Rabitz, Similarity transformation approach to identifiability analysis of nonlinear compartmental models, Math. Biosci. 93 (1989) 217–248.
[26] L. Ljung, T. Glad, On global identifiability of arbitrary model parameterizations, Automatica 30 (1994) 265–276.
[27] X. Xia, C.H. Moog, Identifiability of nonlinear systems with application to HIV/AIDS models, IEEE Trans. Autom. Control 48 (2) (2003) 330–336.
[28] B. Hochwald, A. Nehorai, On identifiability and information-regularity in parameterized normal distributions, Circuit Syst. Signal Process 16 (1) (1997) 83–89.
[29] E.M. Martin, F. Quintana, Consistency and identifiability revisited, Brazilian J. Probab. Stat. 16 (2002) 99–106.
[30] A.M. Chen, H. Lu, R. Hecht-Nielsen, On the geometry of feed-forward neural network error surfaces, Neural Comput. 5 (6) (1993) 910–927.
[31] H. Miao, X. Xia, A.S. Perelson, H. Wu, On identifiability of nonlinear ODE models and applications in viral dynamics, SIAM Rev. 53 (2011) 3–39.
[32] M.P. Saccomani, Some results on parameter estimation of nonlinear systems, Cardiovasc. Eng. 4 (1) (2004) 95–102.
[33] A. Dasgupta, S.G. Self, S.D. Gupta, Nonidentifiable parametric probability models and reparameterization, J. Stat. Plann. Inference 137 (2007) 3380–3393.
[34] C.H. Edwards, D.E. Penney, Differential Equations and Boundary Valued Problems: Computing and Modeling, Fourth Ed., Pearson, 2007.
[35] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, Berlin, 2006.
[36] S.I. Amari, H. Park, T. Ozeki, Geometrical singularities in the neuromanifold of multilayer perceptions, Neural Inf. Process. Syst. 2002.
[37] S. Watanabe, Algebraic geometry of singular learning machines and symmetry of generalization and training errors, Neurocomputing 67 (2005) 198–213.
[38] T.M. Cover, J.A. Thomas, Element of Information Theory, Second ed., Wiley, Chichester, 1991.
[39] R. Bowden, The theory of parametric estimation, Econometrica 41 (1973) 1069–1074.
[40] E.L. Lehmann, Theory of point estimation, Springer-Verlag, New York, 1983.

[41] K. Glover, J.C. Willems, Parameterizations of linear dynamical systems: canonical forms and identifiability, IEEE Trans. Autom. Control 19 (6) (1974) 640–646.

Zhi-Yong Ran received his M.Sc. degree in Applied Mathematics from the Beijing University of Technology, Beijing, China, in 2007. Currently he is a Ph.D. candidate at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include parameter identifiability theory and machine learning.

Bao-Gang Hu (M'94-SM'99) received the M.Sc. degree from the University of Science and Technology, Beijing, China, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, both in mechanical engineering, in 1983 and 1993, respectively. He was a Research Engineer and Senior Research Engineer at C-CORE, Memorial University of Newfoundland, St. John's, NF, Canada, from 1994 to 1997. From 2000 to 2005, he was the Chinese Director of computer science, control, and applied mathematics with the Chinese-French Joint Laboratory, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor at NLPR. His current research interests include pattern recognition and plate growth modeling.